CLAIMS

What is claimed is:

1. A method of automatically creating a dictionary for clustering text documents comprising:

determining a frequency of each word in each of said documents;

creating a Hashtable of most frequently occuring words in said documents;

determining a frequency of phrases in each of said documents that contain only words in said Hashtable;

adding most frequently occuring phrases to said Hashtable; and

outputting said most frequently occuring words and said most frequently occuring phrases as said dictionary.

2. The method in claim 1, wherein said determining a frequency of each word comprises:

removing punctuation and case from said documents;

removing stop words from said document;

replacing words in said documents with synonyms;

removing duplicate words from said documents;

adding remaining words to said Hashtable;

determining said frequency of each word remaining in said Hashtable; and

removing words below a frequency level from said Hashtable.

3. The method in claim 2, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.

4. The method in claim 1, wherein said determining a frequency of phrases comprises:

removing punctuation and case from said documents;

removing stop words from said document;

replacing words in said documents with synonyms;

adding said phrases in each of said documents that contain only words in said Hashtable to said Hashtable;

determining said frequency of said phrases remaining in said Hashtable; and

removing phrases below a frequency level from said Hashtable.

5. The method in claim 4, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.

6. A method of automatically creating a dictionary for clustering text documents comprising:

performing a first pass for each of said documents comprising:

determining a frequency of each word in each of said documents; and

creating a Hashtable of most frequently occuring words in said documents;

performing a second pass for each of said documents comprising:

determining a frequency of phrases in each of said documents that contain only words in said Hashtable; and

adding most frequently occuring phrases to said Hashtable; and

outputting said most frequently occuring words and said most frequently occuring phrases as said dictionary.

7. The method in claim 6, wherein said determining a frequency of each word comprises:

removing punctuation and case from said documents;

removing stop words from said document;

replacing words in said documents with synonyms;

removing duplicate words from said documents;

adding remaining words to said Hashtable;

determining said frequency of each word remaining in said Hashtable; and

removing words below a frequency level from said Hashtable.

8. The method in claim 7, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.

9. The method in claim 6, wherein said determining a frequency of phrases comprises:

removing punctuation and case from said documents;

removing stop words from said document;

replacing words in said documents with synonyms;

adding said phrases in each of said documents that contain only words in said Hashtable to said Hashtable;

determining said frequency of said phrases remaining in said Hashtable; and

removing phrases below a frequency level from said Hashtable.

10. The method in claim 9, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.

11. A program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform a method of automatically creating a dictionary for clustering text documents, said method comprising:

determining a frequency of each word in each of said documents;

creating a Hashtable of most frequently occuring words in said documents;

determining a frequency of phrases in each of said documents that contain only words in said Hashtable;

adding most frequently occuring phrases to said Hashtable; and

outputting said most frequently occuring words and said most frequently occuring phrases as said dictionary.

12. A program storage device as in claim 11, wherein said determining a frequency of each word comprises:

removing punctuation and case from said documents;

removing stop words from said document;

replacing words in said documents with synonyms;

removing duplicate words from said documents;

adding remaining words to said Hashtable;

determining said frequency of each word remaining in said Hashtable; and

removing words below a frequency level from said Hashtable.

13. A program storage device as in claim 12, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.

14. A program storage device as in claim 11, wherein said determining a frequency of phrases comprises:

removing punctuation and case from said documents;

removing stop words from said document;

replacing words in said documents with synonyms;

adding said phrases in each of said documents that contain only words in said Hashtable to said Hashtable;

determining said frequency of said phrases remaining in said Hashtable; and

removing phrases below a frequency level from said Hashtable.

15. A program storage device as in claim 14, further comprising inputting said stop words.

16. A program storage device as in claim 14, further comprising inputting said synonyms.

17. A program storage device as in claim 14, further comprising inputting said frequency level.